A renaissance in library metadata? The importance of community collaboration in a digital world

Bull, Sarah; Quimby, Amanda

DOI:
10.1629/uksg.302

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Bull, S & Quimby, A 2016, 'A renaissance in library metadata? The importance of community collaboration in a digital world', *Insights the UKSG journal*, vol. 29, no. 2, pp. 146-153. https://doi.org/10.1629/uksg.302

A renaissance in library metadata? The importance of community collaboration in a digital world

This article summarizes a presentation given by Sarah Bull as part of the Association of Learned and Professional Society Publishers (ALPSP) seminar 'Setting the Standard' in November 2015. Representing the library community at the wide-ranging seminar, Sarah was tasked with making the topic of library metadata an engaging and informative one for a largely publisher audience. With help from co-author Amanda Quimby, this article is an attempt to achieve the same aim! It covers the importance of library metadata and standards in the supply chain and also reflects on the role of the community in successful standards development and maintenance. Special emphasis is given to the importance of quality in e-book metadata and the need for publisher and library collaboration to improve discovery, usage and the student experience. The article details the University of Birmingham experience of e-book metadata from a workflow perspective to highlight the complex integration issues which remain between content procurement and discovery.

Introduction

SARAH BULL

Assistant Director: Collection Management & Development Library Services
University of Birmingham

AMANDA QUIMBY

E-Resources and Serials Team Leader
University of Birmingham

I was privileged to be asked to speak on the topic of library metadata and standards at the ALPSP seminar 'Setting the Standard' in November 2015[1] alongside a range of excellent speakers representing different stakeholders who have experience in standards development and maintenance in our community. What became clear, after reflecting on my extensive notes during the train journey home that day, is that there is just so much to absorb about how we utilize and contribute to standards to improve our organizations and customer experiences.

There is no getting away from it – metadata and standards can be rather a dry topic of discussion. However, without standards, we would not have an effective supply chain, would not be able to benchmark our services and products and would not be able to deliver content to our users successfully. Without standards, how would we identify ourselves as unique authors? How would we link users to the right 'version' of the journal, book, article or chapter – relative to our perspective? How would we assess value and usage across our services? With thousands of transactions passing through our library systems every year, we would not operate efficiently without standards for transactions, bibliographic metadata, content linking, holdings metadata and usage reporting. They oil the cogs in the machine-to-machine transfer of notifications, events and data.

What is meant by 'library metadata' in this day and age? I find it increasingly difficult to discuss standards, identifiers and metrics as specifically 'library'. By and large, we develop standards as a community. Many of the standards and identifiers that libraries rely on for service provision are also hugely important to the business of publishing. Having been involved with both Knowledge Bases And Related Tools (KBART)[2] and COUNTER[3] initiatives and assisted with their advocacy, it becomes clear that standards need to be based on two core principles: utility to me/my organization and ease of adoption. If both boxes are ticked and our customers rely on the improvements that standards development can facilitate, this generally equals wide-scale uptake. This is a concept described by Stephen Pinfield in his thought-provoking presentation at the ALPSP Conference in 2015[4].

'we develop standards as a community'

Initiatives arising from the UKSG community, such as KBART, COUNTER and the Transfer Code of Practice[5], are excellent examples of where all stakeholders are represented in order to achieve a cohesive and transformative proposition for all.

Libraries and metadata

The 'standards jungle' is a term that I use to refer to the sheer scale and complexity of what we are currently experiencing. Having worked in the information industry for 20 years – a notoriously jargon rich and technically complex landscape – I have never experienced such a time of rapid standards development coinciding with the emergence of organizations and services based around them. The graphic in Figure 1 represents a small proportion of this deluge of standards, protocols, formats and schemas. Libraries work with all of them to a greater or lesser extent.

'a time of rapid standards development'

If we talk specifically about library standards, we need to discuss the importance of bibliographic metadata. Librarians have been crafting bibliographic metadata in-house to a high standard for many years, and continue to do so. However, this is a period of significant disruption in the underpinning standards. Much of this disruption is due to the significant growth of e-content and the need for bibliographic metadata to find a place amongst the wider web of linked data. MARC format metadata continues to be hugely important for content discovery within library search tools, but we also need to consider how we showcase our collections through enterprise and Google search. How do we optimize bibliographic metadata for the web? How does MARC record cataloguing fit with linking to the e-resource dynamically and permanently? The library standards community, IT, repository and library systems organizations are currently trying to answer this question through examining the future role of linked data in bibliographic metadata. There are some potentially transformative projects looking at how bibliographic information can be made more discoverable on the web. Such projects include BIBFRAME[6], the LODLAM community[7] and the Linked Data for Libraries project[8]. Through community engagement, advocacy, training and testing, they are exploring the role of standards in future-proofing bibliographic metadata improvements.

'the need for bibliographic metadata to find a place amongst the wider web of linked data'

However, ensuring developments allow for both current and future cataloguing environments in a rapidly changing standards landscape is really challenging. We need to

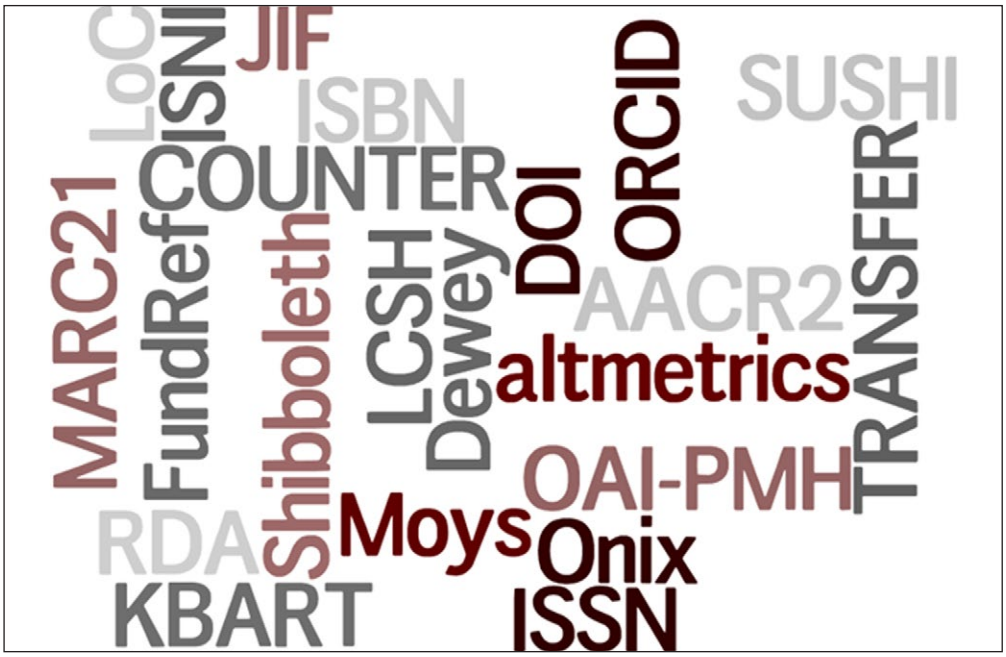

Figure 1. Some of the many standards, protocols, etc. that libraries are required to work with

address issues such as how linked data concepts can be embedded within new catalogue records and also be made backwards-compatible with the wealth of records already available. At present, notwithstanding the projects mentioned above, it is still not clear how the information community transitions its cataloguing systems, standards, practices and integrations to exploit the full potential of web delivery.

My presentation to ALPSP focused more on the here and now. The above disruption aside, libraries still have significant current challenges in ensuring that bibliographic metadata is making visible our unique and distinctive collections in both the library and wider web environment. With increasing reliance on large collections of e-books that may change very dynamically, the age of individual record-by-record manual cataloguing has passed.

Libraries now acknowledge a joint responsibility for e-book metadata with publishers and intermediaries. This has two benefits. Firstly, it gives us the space to explore a more strategic approach to the metadata skillset, by which I mean making use of our expertise in improving our cataloguing and classification policies, focusing on what makes us distinctive in a collections sense and approaching cataloguing from the user perspective through testing and quality control of our discovery services. Secondly, it provides us with the opportunity to work with publishers to facilitate a 'quality' approach to metadata creation. This is beginning to happen.

'Libraries now acknowledge a joint responsibility for e-book metadata with publishers and intermediaries'

Libraries in the past have documented internally, or through organizations such as RLUK[9], a minimum acceptable standard of bibliographic metadata in order to ensure that a record is accurate and ultimately discoverable (through a rich layer of subject and keyword access points) to the end user. Cataloguing is an area which is heavily structured with a wide variety of protocols, controlled vocabularies and formats. At the heart is the MARC record that provides structured fields relating to different metadata elements, specified in fixed positions, which can then be systematically indexed by library catalogues and resource discovery services (RDS). This brings highly relevant results to the user. The fact that they are also library holdings means that there should be no dead ends in discovery – in a print world, at least. If the MARC record provides the means of discovery through 'search', the *classification* of a work in the print world is the 'browse' equivalent – denoting its subject, shelving location and filing order. Subject headings are a means of providing detailed sub-divisions of subject classification which in some cases, such as law and medicine, are highly specialized. In an e-world, location and holdings information is not about shelving order but about the web presence of the work and its constituent parts. MARC format is not geared well to delivering up-to-date, dynamic and durable metadata relating to web location. This is where technologies and standards such as OpenURL, KBART, link resolvers and RDS come in. Bibliographic metadata standards need to integrate with linking and discovery standards which can future-proof (to an extent) against the dynamism of the web. Publishers and content providers also have a big part to play in increasing the stability of references, web pages, citations and references to avoid 'link-rot' (well documented by the Hiberlink Project)[10].

'Bibliographic metadata standards need to integrate with linking and discovery standards'

Working with supplier-derived, standards-based metadata can be a challenge when trying to meet the two criteria mentioned earlier that encourage large-scale uptake of standards, namely, 'utility' and 'ease of adoption'. Publisher organizations sometimes do not have ready expertise and often need help understanding the business imperative. This is where libraries can provide expertise and rationale. Library electronic collections, with an increasing volume of metadata derived outside the library landscape, do not exist in isolation. User discovery is format agnostic around a subject unless specifically tailoring results to online only. Collection strength in a subject discipline is derived from all library holdings – in many formats. This is an underpinning principle of an RDS which offers that context-sensitive, single-search environment. With a decreasing amount of bibliographic metadata under direct library curation, it is important that a discussion happens within the community over good practice in metadata provision to library systems. This is not about controlling an environment per se but

'it is important that a discussion happens within the community over good practice in metadata provision'

about making the most of expensive purchases through improved user discovery. What follows is a specific case study at the University of Birmingham where we are in the process of changing our approach. [Over to Amanda.]

The University of Birmingham approach

Over five years ago, like many institutions, we had two separate catalogue interfaces: an OPAC for our print collections and an e-library for our electronic collections via a link resolver and federated search system. We had an established e-journal collection and were at the early stages of developing a significant e-book collection. With e-books being acquired in collections and with titles being frequently added and removed, it became impossible to catalogue each e-book individually. We needed to streamline our metadata processes. Due to limitations in our original library systems and internal infrastructure, we could not efficiently import e-book MARC records in bulk. So, we abandoned MARC records and relied instead on the link resolver to route users to e-book content.

'it became impossible to catalogue each e-book individually'

In 2012 we launched Primo, our RDS. As a result, all our print and electronic library collections were brought together into a single discovery layer with increased granularity provided by journal articles and e-books/chapters via a pre-harvested central index. Although a very positive development, combining multiple underlying data sources and incorporating the vast centralized index has significantly increased complexity regarding relevancy ranking. It is even more difficult to ensure that the most appropriate content is discoverable.

Our e-book collection has almost doubled in the last five years to over 500,000 titles (compared with just over two million print books). We use a variety of different business models on a spectrum from 'just-in-case' to 'just-in-time' access. With our experimentation of patron-driven acquisition (PDA) and evidence-based acquisition (EBA), it became essential to review our MARC record decision again. This was reinforced due to an initial low usage of our PDA profile.

'it became essential to review our MARC record decision again'

A link resolver has never been designed as a discovery source, rather a means of linking. Developments in pre-harvested central indexes such as PCI (Primo Central Index) are replacing poor link resolver metadata records with richer, more comprehensive metadata. The link resolver then provides the onward dynamic OpenURL link.

Based on our previous experiences, one condition that a move back to MARC record import had to fulfil was that it be sustainable for librarians to manage the process. We have not managed to use our library management system (LMS), Aleph, as a data source for e-books due to historic problems with poor e-book records which are currently suppressed. In order to begin using MARC records again, the Digital Library Team within IT Services did the initial groundwork of creating an FTP site to hold MARC files from suppliers/publishers and developing a publishing workflow. This metadata is normalized and enhanced by Library staff before being uploaded to the site using the open source Filezilla[11] FTP client. The site is parsed via a Linux shell script on a daily schedule to produce an import file of added or changed MARC files which is ingested into Primo using a customized and localized version of ExLibris' MARC import normalization rules. Although the MARC records being ingested contain URLs that link directly to that supplier's version of the title, our 'View online' tab within the RDS record instead works by sending the ISBN to our link resolver, SFX. Through the display rules of our link resolver, users are then presented with owned copies in preference to the PDA versions. This is necessary because although we have de-duplicated our PDA profile with our owned collections, as a safeguard we hide the PDA version from displaying if any de-duplication has been missed.

The Collection Development Team in Library Services was then able to handle the management of the MARC records and link resolver activations. We did this initially just with the PDA MARC records provided by our supplier to test the process. This process involved:

New content:

- obtaining MARC records from the suppliers during the acquisition process
- checking the supplier MARC file against the corresponding link resolver target to ensure a complete content match
- activating titles on the link resolver
- quality checking MARC records
- adding the MARC file to the FTP site for indexing in Primo.

Updates to our collection:

- activating on the link resolver before a MARC record is added, otherwise you end up with a MARC record and no link.

Deletions:

- deleting the MARC records first because the link resolver record is suppressed from view without a MARC record.

As a result of the new workflow, we have a quality MARC record from our FTP site directly displayed in our RDS, with the link resolver layer underneath for linking.

'we have a quality MARC record …with the link resolver layer underneath for linking'

During the testing and import process, we learned a lot about the quality of the supplier-derived records. We found that although the MARC records were considered better than the link resolver records, they were still far from an acceptable standard in some cases. Therefore as an additional step, our Metadata Team also learned how to use the MARCedit[12] tool to bulk edit the MARC files to bring them up to an acceptable standard. Figure 2 describes some of the problems we encountered.

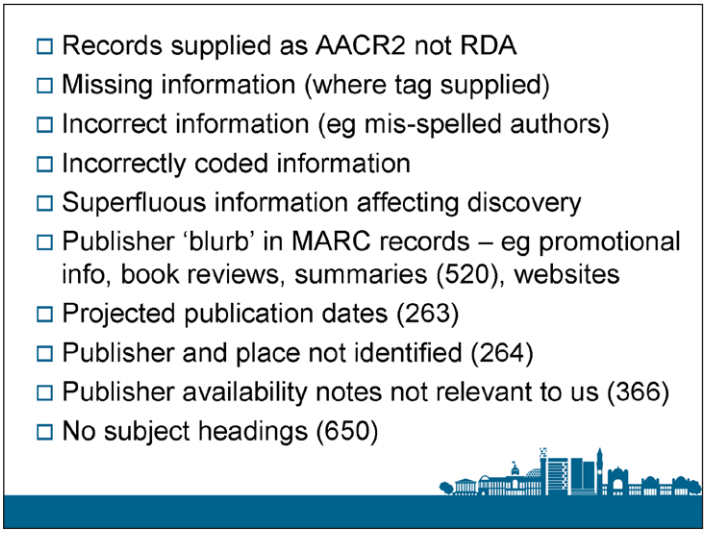

Figure 2. Some of the problems encountered when working with MARC records

The impact

Within the first two weeks of adding e-book MARC records to our RDS for our PDA profile, our overall usage almost doubled compared to the previous two weeks (as shown in Figure 3).

'our overall usage almost doubled'

Introducing MARC records for PDA e-books improved visibility in the results screen. With expenditure rising, it became evident that we urgently needed to treat our owned collections in the same manner. We therefore requested, imported, reviewed, edited and activated many further collections of e-books. We now have approximately 140,500 MARC records available through the FTP site and continue to retrospectively source catalogue records for the remainder.

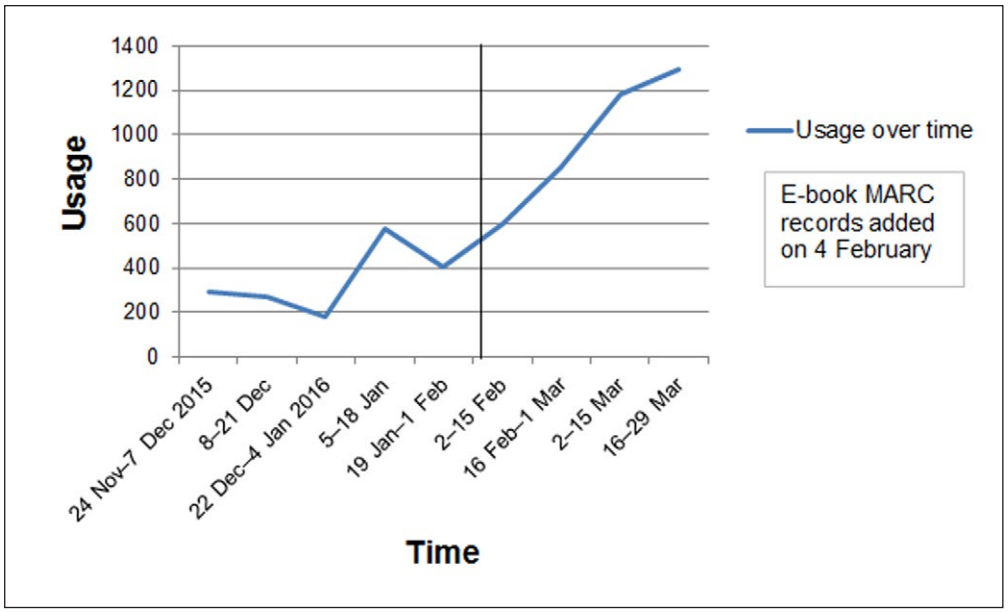

Figure 3. Increased usage after adding e-book MARC records to the RDS for their PDA profile at the University of Birmingham

Issues

1. Quality metadata within MARC records

Even though the use of MARC records is an improvement, we have invested additional resource to improve the MARC records that we have obtained from suppliers, notably aggregators, to achieve an acceptable standard. Some suppliers are now improving quality but there needs to be communication and collaboration with the library community on requirements.

2. E-book collections within PCI

To maximize the discoverability of e-books, we enabled supplier catalogues within PCI so that they were merged with local data sources. While this can produce search results for titles we have not acquired – on platforms where we do not subscribe to 100% of the content – the use of our link resolver directs users to alternative, owned content where available. The downside is that multiple instances of the same e-book appear in search results and are not de-duplicated. Also, the ranking algorithm used will often place our MARC record below the PCI equivalent – a situation that is under ongoing investigation. We need to better understand the value of PCI e-book records alongside a local MARC record with OpenURL link.

3. Delays in discovery index updates

The process of deleting from dynamic e-book collections can be handled by deactivating e-books on the link resolver and also adding a delete MARC file to our FTP site. However, a PCI record can take between three to 14 days to disappear from results. Thus there is a period of time when there is no OpenURL link under a PCI record resulting in a dead end for the user.

4. Matching MARC file lists with a link resolver knowledge base

ISBN metadata within a supplier's MARC file does not always match the link resolver ISBN metadata (used for creating the OpenURL link). A mismatch means that the relevant link resolver object is not activated and there is no link resolver layer underneath the MARC record.

Our experience at the University of Birmingham and the issues discussed above really underscore the fact that this is very much still an experimentation of workflow within the Library and with the wider community. We need to understand more about the value of the pre-harvested central index approach to surfacing e-book content in relation to both MARC records and linking. However, in general it has made us appreciate the work still to be done to make metadata for e-books an efficient part of the acquisition and discovery process. [Back to Sarah.]

'still an experimentation of workflow within the Library and with the wider community'

Conclusion

As you can see, with the piloting of new dynamic access models for e-books alongside our need to improve the methods by which we make such content discoverable, we have seen intensively some of the issues with metadata that can arise from a lack of standard processes or integration. As a community there are some key improvements that we need to have conversations about in order to improve our collective awareness. Some questions which might form part of this conversation are as follows:

- Can we reach consensus on good practice for e-book bibliographic metadata generated by suppliers and publishers?
- Can we agree, as part of our supply chain relationships, targets for timely, accurate and comprehensive bibliographic and knowledge-base metadata to avoid delays in discovery and usage?
- Given that an increasing percentage of our library catalogues contain metadata that is not internally generated, can we reassess the importance of metadata supply?
- Can we discuss the role of e-book records in central discovery indexes including more timely updates to indexes?
- Can we present a clear business case for engagement with metadata that speaks to all individual stakeholders – around discovery, usage, best value, impact, and relationship management? These should be central values to all of us regardless of whether we are a cost centre or a revenue business.

Some of these questions are already being discussed in standards initiatives like ODI[13] and KBART but they need to achieve greater community engagement and uptake.

Development of standards and best practice is at the core of our community and central to the success of an effective supply chain. We all use standards, either as providers, intermediaries or consumers, but do we also contribute to their refinement, adoption, improvement and governance? It is so important for us to participate rather than just utilize and consume. Otherwise, we end up with institutional data silos, poor interoperability and proprietary solutions that fail to connect our world. We all have a role in ensuring that standards are timely, work across the community, enable ease of uptake and are constantly improved through community input. Yes, standards are not the most exciting area of activity but, without them, the supply chain will not work effectively and our customers' experiences will be the poorer.

'It is so important for us to participate rather than just utilize and consume'

Abbreviations and Acronyms

A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the 'Abbreviations and Acronyms' link at the top of the page it directs you to: http://www.uksg.org/publications#aa

Competing Interests

The authors have declared no competing interests.

References

1. ALPSP:
   http://www.alpsp.org/Past-Events/Setting-the-Standard/30461 (accessed 28 April 2016).
2. NISO:
   http://www.niso.org/workrooms/kbart (accessed 28 April 2016).
3. COUNTER:
   http://www.projectcounter.org/index.html (accessed 28 April 2016).
4. Pinfield, S:
   https://www.youtube.com/watch?v=Bgx9IQdXdVA (accessed 28 April 2016).
5. NISO:
   http://www.niso.org/workrooms/transfer (accessed 28 April 2016).
6. BIBFRAME:
   https://www.loc.gov/bibframe (accessed 28 May 2016).
7. LODLAM Community:
   http://lodlam.net (accessed 28 May 2016).
8. Linked Data For Libraries (LD4L):
   https://www.ld4l.org (accessed 28 May 2016).
9. Research Libraries UK (RLUK):
   http://www.rluk.ac.uk (accessed 28 April 2016).
10. Hiberlink:
    http://hiberlink.org (accessed 28 April 2016).
11. Filezilla:
    https://filezilla-project.org (accessed 28 April 2016).
12. MarcEdit Development:
    http://marcedit.reeset.net/about-marcedit (accessed 28 April 2016).
13. NISO:
    http://www.niso.org/workrooms/odi (accessed 28 April 2016).

Sarah Bull

Assistant Director: Collection Management & Development Library Services
University of Birmingham, Edgbaston Birmingham B15 2TT, UK
Tel: +44 (0)121 414 7117 | E-mail: s.price@bham.ac.uk

ORCID ID: http://orcid.org/0000-0002-0484-0272

Amanda Quimby

ORCID ID: http://orcid.org/0000-0002-3785-6254

To cite this article:
Bull, S L and Quimby, A, A renaissance in library metadata? The importance of community collaboration in a digital world, *Insights*, 2016, 29(2), 146–153; DOI: http://dx.doi.org/10.1629/uksg.302

Published by UKSG in association with Ubiquity Press on 05 July 2016